DOCUMENT RESUME

ED 356 558 EA 024 821

AUTHOR Nye, B. A.; And Others
TITLE Class-Size Research from Experiment to Field Study to Policy Application.
PUB DATE Apr 93
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (74th; Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Class Size; Elementary School Students; Longitudinal Studies; Primary Education; Public Schools; *Student Development; Teacher Role; *Teacher Student Ratio
IDENTIFIERS *Lasting Benefits Study TN; *Student Teacher Achievement Ratio Project TN

ABSTRACT

A 4-year longitudinal study in Tennessee, called the Student-Teacher Achievement Ratio (STAR) Project, examined the effects of class size on student achievement in kindergarten through grade 3. More than 6,000 students from 75 schools in 42 school systems were included in the study. There were three class sizes: small class (13-17 students), regular class (22-26 students), and regular class with a full-time teacher aid. The study found that students in small classes improved more than students in larger classes. Gains children made in kindergarten were maintained through grade 3. Analyses show that class size had an effect in all locations. However, the presence of a teacher aid did not show an effect. The Lasting Benefits Study (LBS) followed STAR students through grade 4 and grade 5 to determine the lasting effects of early small-class involvement. Students who were in STAR small classes in grade 3 were more advanced statistically and educationally and had higher school participation measures than students who were in regular classes. Another study, Project Challenge, provided incentives for class-size reduction in 17 Tennessee counties (1990-1992). Preliminary results show small-class students gained in reading and mathematics levels. Tables and appendices on data-collection methods and results are included. (Contains 32 references.) (Author/JPT)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED356558

CLASS-SIZE RESEARCH
FROM EXPERIMENT TO FIELD STUDY TO POLICY APPLICATION*

A Report Incorporating Three Class-Size Initiatives:
Tennessee's Student Teacher Achievement Ratio (STAR) Project (8/85-8/89),
Lasting Benefits Study (LBS: 9/89-7/93), and Project CHALLENGE (7/89-7/93)
as a Policy Application (Preliminary Results)

Authors:

B.A. Nye
C.M. Achilles
J.B. Zaharias
B.D. Fulton

AERA
April, 1993
Atlanta, Georgia

U.S. DEPARTMENT OF EDUCATION
Office of Educational Re[illegible]h and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. M. Achilles

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

* The authors acknowledge the contributions of the entire Student Teacher Achievement Ratio (STAR) Project staff, especially to E. Word, Tennessee State Department of Education, Project Director; H. Bain, J. Folger, J. Johnston, and N. Lintz who were the other members of the STAR Consortium; J. Finn, R. Hooper, and G. Bobbett, Consultants.

2/93

BEST COPY AVAILABLE

EA 024 821

ERIC

CLASS-SIZE RESEARCH
FROM EXPERIMENT TO FIELD STUDY TO POLICY APPLICATION*

A Report Incorporating Three Class-Size Initiatives:
Tennessee's Student Teacher Achievement Ratio (STAR) Project (8/85-8/89),
Lasting Benefits Study (LBS: 9/89-7/93), and Project Challenge(7/89-7/93)
as a Policy Application (Preliminary Results)

B.A. Nye, C.M. Achilles, J. Boyd-Zaharias, B.D. Fulton**

ABSTRACT

This paper describes processes and results of three related class-size studies that move through three stages: experiment, field study and policy application. They constitute a major longitudinal contribution to education research.

Education leaders in Tennessee supported a four-year (8/85-8/89) longitudinal study of class-size effects on pupil achievement in early primary grades (K-3). The project included over 6000 pupils/year in 75 schools in 42 school systems. There were three experimental conditions: Small class (13-17), Regular class (22-26) and Regular class with full-time teacher aide. Pupils were randomly assigned to class-size conditions; teachers were randomly assigned to classes. Pupils in small classes (1:15) made significantly (statistically and educationally) greater gains than other pupils, and minority pupils in small classes benefitted more than minority pupils in other class conditions. Gains initiated in kindergarten were maintained through third grade. Analyses showed a continuing, powerful class-size effect in all locations. There was no consistent teacher-aide effect evident in the analysis. This large-scale randomized experiment provided some definitive answers about class-size effects in early primary grades. The LBS has already followed a sample (n=4320) of STAR pupils through grades 4 and 5 (1989-92) to show the lasting benefits of early small-class involvement. In LBS students who were in STAR small classes in grade 3 are statistically and educationally ahead of students from Regular and Regular/Aide STAR class conditions. The small-class students also have advantages in school participation measures. Project Challenge provided incentives for class-size reductions in 17 of Tennessee counties (1990-1992). Preliminary results show evidence of pupil gains in reading and math in Challenge.

---

* The authors acknowledge the contributions of the entire Student Teacher Achievement Ratio (STAR) Project staff, especially to E. Word, Tennessee State Department of Education, Project Director; H. Bain, J. Folger, J. Johnston, and N. Lintz who were the other members of the STAR Consortium; J. Finn, R. Hooper, and G. Bobbett, Consultants.

** B.A. Nye, Director, Lasting Benefits Study and Center of Excellence for Research in Basic Skills, Tennessee State University, Nashville, TN 37203. C.M. Achilles, Professor, Education Administration, UNC-Greensboro, Greensboro, NC 27412-5001. Member of the Star Consortium 8/85-9/88, and consultant to LBS, 89-92. J. Boyd-Zaharias, and B.D. Fulton, Staff of LBS at the Center for Excellence.

FROM EXPERIMENT TO FIELD STUDY TO POLICY APPLICATION

Introduction: Some Critique of Class-Size Issues and Longitudinal Results

Education researchers seldom conduct either experimental or longitudinal study. Less often do researchers apply and study results of experimental and longitudinal research. Education research does not often provide clear direction for education practice. In contrast, this paper discussed a continuing strand of research that 1) began in 1985 as experimental and longitudinal (through 1989), 2) is still using and extending the original data base (1989-1992), 3) has provided policy direction and implementation (1989-1992), and 4) continues to spawn a variety of interesting ancillary studies.

Some things make so much sense that people wonder why researchers study them. Class size -- the number of pupils that a teacher works with at a given time -- is one such issue. Early studies were usually short-term, poorly designed, and dealt with reductions in large units (say 45-30 pupils). A meta-analysis (Glass & Smith, 1978) and critiques of it (Education Research Service, or ERS, 1978 and 1980) heated up the debate. Continuing policy discussions (Glass et al., 1982: Cahen et al., 1983) encouraged Tennessee legislators to commission a large-scale, longitudinal experiment of class size issues. While Tennessee's Student/Teacher Achievement Ratio (STAR) study was on-going, policy debates continued (e.g., Mueller et al., 1988; Tomlinson, 1988; Mitchell et al., 1989).

After STAR results became public (Word et al., 1990), some collections of works on class size reviewed the findings and ideas related to policy (e.g., Robinson, 1990; Contemporary Education, 1990: Peabody Journal of Education, J. Folger (Ed.), 1989, published in 1992). The Robinson (1990) report did not yet have complete details from STAR, but did say, "Tennessee's Project STAR, currently in progress. . .had positive effects as measured by scores on nationally standardized tests (grades K-2)" (p. 82). Other studies reported generally

positive results for STAR and mixed results for other "class size" studies. STAR has had some critics. Response to some STAR criticism offers insight into the issues.

Recent policy discussions (e.g., Tomlinson, 1988 and 1990; Mitchell et al., 1989; 1989/92) seem to take views that 1) small class size is expensive, 2) there may be more efficient but equally good early interventions, 3) teachers only want smaller classes to have less work, etc. The analysts don't provide data to refute class size gains found in the few well controlled studies. In attempting to hew tightly to conservative administration policies, federal- employee Tomlinson (1990) blends both absolutely incorrect information and a mixture of praise and pejoratives in discussing STAR. The following are examples (Tomlinson, 1990, p. 19). Comments on the quotes are in parenthesis

> Project Star has indisputably shown us that for a period of one year, classes that averaged 15 children learned more than classes that averaged 23. (True, but. . .the children learned more each year for four years, and researchers are still studying the "Lasting Benefits" of small classes in STAR.) (Word et al., 1990)
>
> Project STAR is doubtless the all time most comprehensive controlled examination of the thesis that a substantial reduction in class size will, of itself, improve achievement. (True. A praise.)
>
> It will doubtless remain in a class by itself because of the inherently impractical cost of the research and its putative implications for class size, the uninteresting theoretical implications of the findings, and, yes, the uncertainty that still remains about the causes (emphasis added) of the observed improvement. (Pejorative praise? The design clearly leaves little doubt about the findings and causes.)(Word et al., 1990)
>
> The principal finding of Project STAR. . .is the mundane substantiation of a class size effect. . . . (This is a strange comment in a field where the research is usually denigrated for not finding effects.)
>
> Perhaps more worrisome was the fact that a significant class size difference was found only in the first year of the three-grade study (plus kindergarten). (Absolutely not true statement.) (Finn et al., 1990; Word et al., 1990)
>
> Teachers volunteered to participate. (Absolutely not true statement. Teachers were randomly assigned all four years.) (Word et al., 1990)

There have been some unambiguous positive statements made about STAR. The Orlich (1991) statement is gratifying: ". . .in my own opinion, (STAR) is the most significant educational research done in the US during the past 25 years" (p. 632). Two may be downplayed slightly as they were made by STAR researchers (but reviewers of major journals recommended publication). "This experiment yields unambiguous evidence of a significant class size effect, at least in the primary years" (Finn et al., 1990, p. 135), and "This research leaves no doubt that small classes have an advantage over larger classes in reading and mathematics in the early primary grades" (Finn & Achilles, 1990, p. 573).

Perhaps the most confusing criticism is the one offered by Mitchell et al. (1989/1992) who review test results (after the intervention) and state, "For some reason, low performing students are more often found in larger classes while their high performing counterparts are about equally distributed between large and small class settings, reducing the achievement level of regular classes while raising that in the smaller ones" (pp. 65-66). The critics use the experiment's results to question its strength and, finding that STAR achieved the class size effect, Mitchell et al. (pp. 63, 66, 67) suggest "non-randomness" ("Either the parents. . .were able to influence the placement. . ."; "combined with the peculiarities of student assignment. . ."; "combined with unexplained non-randomness. . ."; etc.). Their discussion of "non-random" is based on the testing results at K. That is, they use the effects (what the study showed) to try to explain non-randomness!

Rather than look at test results one year after the intervention, they could have checked available data that were not connected with effects. These data could be demographics. The STAR researchers did check demographics of districts/schools in the random sample and found "no differences" except in district pupil enrollments where inclusion of systems in the four largest-population counties in Tennessee caused STAR districts to have a slightly larger average pupil enrollment than non-STAR districts (Word et al., 1990). Another "randomness" check would be to review the proportion/percent of pupils with certain demographic characteristics against the proportion/percent of pupils in the three class conditions.

Proportions of students (sex, race, free lunch, special ed) in each class type (Small or S, Regular or R, Regular with Aide or RA) when compared to the total distribution of students by class type shows a "random" picture (Table 1). One exception is in special ed where S classes had a high proportion of identified special ed pupils. For example, S classes included 30% (n=1900) of the 6325 STAR pupils in K. In the STAR sample, 30.1% of the males, 30.0% of the females, 29.0% of the non-white, and 30.6% of the white (etc.) pupils were in S classes.

Table 1 about here

What is it about STAR (and its on-going derivatives the Lasting Benefits Study or LBS and Project Challenge) that seems to generate strong positions, even among those who, at least as suggested by their roles in education, should support research that shows ways to benefit pupils? Research should be subject to serious peer review and critique -- especially research identifying expensive options or research that seems to provide expert verification of practitioner and common-sense wisdom -- but the reviews should be accurate, scholarly and without innuendo or "cheap shots." Let's review the studies.

PHASE I. STAR: THE BASIC STUDY AND DATABASE: DESIGN AND SCOPE

Project STAR began in 1985 with pupils in Kindergarten (K). All Tennessee districts were asked to participate. Due to the scope of the study, researchers (using a "power analysis") determined that they would need approximately 100 classes of each of three class types (S with average 1:15 teacher/pupil ratio -- range 1:13-1:17; R with 1:24 average -- 1:22-1:26 range, and RA with 1:24 average and a full-time Aide). Forty-two of the 140 districts (1985) were selected, and 79 elementary schools in those districts (voluntarily) provided the sites for STAR intervention. Three districts eventually dropped out.

Sites had to agree to participate for four years, to have some visitations and extra testing, and to allow random assignment of pupils and teachers to conditions. Sites had to have space for the added classes and at least 57 pupils in K. This did exclude very small schools from the study, but at least 57 pupils were needed for the in-school design (minimum of 1:13, 1:22,

1:22) that assured that any school with the S class also included R and RA class conditions. This powerful design helped ameliorate building-level variables such as leadership, curriculum, facilities, expenditures, SES, etc.

The state paid for additional teachers and aides for the four-year study (K-3) from 1985-1989. The STAR study made only class-size changes. Districts followed their own policies, curricula, etc. No pupil in STAR would receive less (e.g., would have a disadvantage from the state norm) by being in STAR. Not every pupil took every test or had every data point, so for a given year the n for analysis was less than the total of pupils participating for that year. (Table 2 shows that 5734 of the 6325 K pupils provided the K analysis group.) All pupils in an analysis had all data needed for that analysis.

Table 2 about here

STAR employees monitored testing conditions for consistency. Although the pupil was the primary unit of data collection (researchers collected teacher, principal, district data and such things as teacher interviews, etc. to support the class size analysis), the class was the unit of analysis (It was a study of class size effects.) This analysis recognized that each pupil is not an independent measure -- the teacher and classmates all influence the learning environment.

Legislation required that STAR classes be in four locations: inner city, urban, suburban and rural. The major question was: "What is the effect of reduced class size (e.g., 1:15) on pupil achievement and development in K-3?" Research was conducted by a consortium of four universities, each with a principal investigator and staff (University of Tennessee, Memphis State, Tennessee State, and Vanderbilt) and the Tennessee State Education Agency (SEA) where the director was housed. Persons from each university monitored the study in assigned schools. (Ancillary studies reviewed training effects, teacher/teaching practices, etc.) This report primarily reviews achievement.

Achievement was determined by pupil scores on both Norm-Referenced Tests (NRT) and Criterion-Referenced Tests (CRT) appropriate for the grades. The CRT was Tennessee's Basic Skills First (BSF) test tied to the state curricula. (Appendix A is a list of data measures.)

Due to the randomness the basic design was post-test only (pre-test in K was not an option). With scaled scores it was possible to study year-to-year gains as STAR tracked each pupil and as pupils were in the same class size condition from year to year. When pupils moved to/from STAR schools, replacement was random.

STAR Design/Analysis/Selected Findings*

The general multivariate design included four locations and the class type (S, R, RA) for either achievement measures or non-cognitive measures. The design also included pupil (and teacher) characteristics of interest, and in grade 2, issues of teacher training. The primary analyses addressed the required questions as stated in the legislation and were completed for each of the four years. Additional longitudinal analyses are underway. (Details are available in STAR technical reports from the STAR office, Tennessee SEA, Cordell Hull Building, Nashville, TN 37219.) The outline for the primary analysis and the extended model for the detailed analyses are in Appendix B. The primary analysis consisted of multivariate tests of mean differences between and among the groups being analyzed. [This design is also being followed in the Lasting Benefits Study (or LBS) effort to the degree possible.]

The analysis employed a general linear model approach for unequal-n design. The design has unequal n's and some empty cells and requires multiple error terms to test all of the fixed effects. Test statistics were the univariate F-ratio for each measure and Wilks' likelihood ratio

* The STAR Consortium used an external advisory board and an external consultant to conduct independent analyses of STAR data. Project and external analyses were confirmatory. The achievement analysis involved Stanford Achievement Tests, or SAT, and Tennessee's criterion-referenced BSF tests. The Consortium chose SESAT II over SESAT I since Tennessee (K) objectives correlated better with SESAT II than with SESAT I, and SESAT II offered a higher "ceiling," allowing pupils to show greater gain. The Consortium also chose "comparison" schools selected from STAR districts which already used the SESAT II, SAT and other tests. Analyses of STAR results with comparison-school results have yet to be done.

for multivariate sets. Other analyses and tests (e.g., chi square, correlation, regression) were employed as needed. There were two planned contrasts tested among three class types:

S class mean vs. all R and RA class means (S vs. "Other")
R Class mean vs. RA class mean

The major achievement results of STAR appear in Appendix C. (For STAR, development measures such as attendance, discipline and self-concept showed no differences between S and R/RA.) In many ways the monotony of the findings is significant. Essentially, pupils in S did statistically significantly better (usually at $p \leq .001$) than pupils in R and/or RA. The class size effect was found equally in all locations (e.g., urban, rural) and favored the S condition in all four grade levels. Some less pervasive findings appeared in single grades, or in two of the four years.

Some simple analyses demonstrated powerful effects. Note (Table 3) that in the average percent of pupils passing the CRT (BSF) in grade 1 there appears to be a strong positive class size benefit for minority pupils (This result was confirmed in more "sophisticated" analyses but the results in Table 3 speak for themselves.) Over 17% more minority pupils pass the BSF if the pupils are in S rather than in R (or RA). The gap for minority students gets insurmountable in grade one in large classes, and remediation (e.g., Chapter 1) has not seemed to close the gap in the past. This gets expensive.

Table 3 about here

The statistical significance question seems to be resolved in class size issues. There remains the "educational" significance question. Often "educational" significance is dealt with by reviewing the "effect sizes." Effect size is one way to see how much the gain is relative to a standard deviation. With the CRT the educational effect might be the percent passing, as percent has a standard of 100. Effect sizes favoring S in STAR range from .08 (in K) to .40 (in grade 3) for minority pupils. Generally the positive STAR effect sizes for pupils in S are in the .20 to .27 range. (See Table 4.)

Table 4 about here

PHASE II. THE LASTING BENEFITS STUDY (LBS)

STAR results are clear. What happens, however, when these pupils who benefitted from S in K-3 return in grades 4 and later to "regular" classes? Weikart (1989) and material in Futurist Magazine ("Education," 1990) point out the lasting benefits of early intervention. The STAR database provides the opportunity for a longitudinal study of benefits of early small-class involvement. The LBS is primarily a process to follow pupils who were in STAR in the S, R, RA conditions. Analyses use pupil test scores and behavioral indicators of school efforts. The fourth-grade analysis included 4230 pupils. (They were identified by class type in at least grade 3.) Of those 1412 were S, 1250 were R and 1568 were RA. [Note: Analyses of grade-5 test scores have provided results similar to grade-four analyses. These are shown in tables with the grade-4 results. Grade-6 results seem to be like grade-5, but are "in process."] The LBS lacks the benefits of the extreme design strengths of STAR; LBS is "field research" while STAR was a true "experiment." Nevertheless, the LBS results are informative.

Scaled-score means for the three STAR class types (S, R, RA) were compared through multivariate analysis of variance (MANOVA) for unequal n's using the MULTIVARIANCE program (Finn & Bock, 1985). The analysis examined mean differences among three class types, the mean differences among four school locations (rural, urban, suburban, inner-city), and the interaction between class types and locations. Using the basic STAR analysis design, three achievement subsets for the LBS were compared separately. Two subsets include scores from both the NRT and CRT components of the Tennessee Comprehensive Assessment Program or TCAP. Set 1 included Total Reading (NRT scores), Total Language (NRT scores) and the number of domains mastered in Language Arts (CRT). Set 2 consisted of Total Math (NRT scores), Total Science (NRT scores), and the number of domains mastered in Mathematics (CRT). Set 3 included Study Skills (NRT) and Social Science (NRT) scores. (See also Finn et al., 1989/1992).

The LBS analysis yielded clear and consistent results. Students previously in a small-size STAR class demonstrated in every location that they had statistically significant ($p. \leq .01$) advantages over R and RA pupils on every set of measurements. The greatest achievement advantages were for inner-city and suburban classes (Table 5). For grade 5 all S v R contrasts were significant ($p \leq .01$); no R v RA contrast was significant.

Table 5 about here

The Project STAR results indicated substantial educational benefits for students in small classes. The positive effects from involvement in a small-size class still remain pervasive one full year after students returned to regular-size classes. The LBS students who had attended small STAR classes had an educationally and statistically significant advantage over LBS students who had attended R or RA STAR classes. This advantage can be measured by the TCAP scaled-scores differences between S and R classes, and between the RA and R classes as shown in Table 6. Students from the S classes retained their academic advantage.

Table 6 about here

Table 7 provides estimates of the S and RA class effect sizes, grades 4 and 5, 1989-90 and 1990-91. Effect sizes ranged from .11 to .34 for the S/R contrast. The R/RA contrast shows effect sizes ranging from -.02 to -.09 (Finn et al., 1989/1992; Nye et al., 1991, 1992). The significant advantages for LBS fourth-grade students who had been in STAR small classes form a strong pattern of consistency. Small-class students outperformed R and RA class students on every achievement measure in all locations.

Table 7 about here

As part of the LBS analysis Finn et al. (1989/1992) reported differences in student participation based on prior class-size experiences (S, R, RA). (Details of the participation

idea appear in Finn, 1989 and in Finn & Cox, 1992). Essentially, according to Finn (1989) increased student participation in school reflects a decreasing tendency for student alienation and dropout in later years. To a great extent opportunities for student participation (e.g., clubs, service projects, government, music, athletics) can be established and operated by those in schools -- teachers and administrators. Participation can also include the pupil's active involvement in classroom activity.

Finn et al. assessed a grade four subset of STAR pupils by asking their teachers to rate them on the 25 item Pupil Participation Questionnaire on a five-point range from (1) "never" to (5) "always." Teachers rated pupils on three behavioral scales (Finn et al., 1989/1992).

> . . .Nonparticipatory Behavior (e.g., "Annoys or interferes with peers' work"), Minimally Adequate Effort (e.g., "Pays attention in class"), and Initiative Taking (e.g., "Does more than just the assigned work"). (p. 78)

Teachers rated pupils in their classes who had participated in one of three STAR conditions for three years (grades 1-3). The 258 teachers in 74 schools rated 2,207 pupils. Using the STAR and LBS MANOVA design, scores on the three participation scales -- Effort, Initiative and Nonparticipatory Behavior -- were simultaneous criterion variables (p. 79). Statistically significant differences were found on participation variables:

> [Location ($p \leq .05$); Class type ($p \leq .0001$); Loc x Type ($p \leq .05$)] (p. 79).

According to Finn et al. (1989/1992):

> The particular contrast of small-class with regular-class students was statistically significant at $p \leq .05$ using a multivariate test and at p-values of .05 or .01 on individual scales. Pupils who had attended small classes were rated as having superior modes of participation in grade four in comparison to their peers. (p. 81)

The participation effect sizes (.11 to .14) were similar to effect sizes found in LBS achievement analyses (.11 to .16) The R/RA contrast was not significant. The grade four LBS study shows that the STAR small-class benefit is retained consistently one full year after STAR ended. There is also the added benefit of increased participation behavior -- positive behavior linked to staying in school (Finn, 1989). This LBS analysis links the desired participation

behavior to higher academic achievement on measures used in LBS. (Although not obtained for the grade-five analyses, LBS researchers plan to assess participation again.)

Building upon the database provided by STAR, LBS is showing that early small-class involvement (e.g., 1:15) has continuing benefits (note also Weikart, 1989). This does, in effect, deflect some criticism of the cost of reduced class size, since the benefits are spread out over more years than simply during the years of the class-size reduction.

PHASE III. PROJECT CHALLENGE AS POLICY IMPLEMENTATION

To help pupils in selected Tennessee counties , the state provided funding and incentives for local district leaders to use various strategies to improve pupil performance. Beginning in 1989, one option -- called Project Challenge -- was to reduce the class size in 17 districts in grades K-3 to approximately 1:15. Project Challenge put into practice results of the statewide STAR experiment.

Prior to 1989-90 Tennessee pupils took the Stanford Achievement Tests (SAT) as the state testing format. Beginning in 1989-90 students in selected grades began taking the Tennessee Comprehensive Assessment Program or TCAP. The TCAP includes both a NRT and a CRT component. Since no special testing was done for Challenge, extant data and regular testing processes were used in the evaluation plan. Test data and results for all discussions are for grade two, the first grade for regular TCAP testing on a statewide basis.

The Tennessee SEA needed some idea if the class size reduction (1:15) seemed to be helping student achievement in the 17 counties. Since Challenge was not an "experiment" with random selection or assignment, special testing, etc., an evaluation is essentially an after-the-fact (post hoc) review and analysis of grouped (e.g., school system) data, using the available second-grade test results. There is no sure way to attribute any gain (or loss) to Challenge (e.g., class-size reduction) if other special "interventions" were taking place at the same time in the same grades. There may be other systematic threats to validity, too. Grouped data by grade level are subject to any variation in student ability by classes or grades. Gains or losses in one year may be the result of very good (or very poor) student ability, excellent teaching,

test variation, etc. Only with several years of results can a trend become evident. Experience with STAR and LBS can help in Challenge.

Thus, since testing changed in 1989-90 and Challenge began in 1989-90, use of 1989-90 second grade TCAP results as the baseline data for Challenge means that the second-graders in 1989-90 already had one year of Challenge (that is, 1989-90 data are baseline after one year of treatment). Use of 1990 TCAP as "baseline" even when pupils had one year of "treatment" seemed preferable to using the pre-Challenge but not comparable SAT results for second graders. The 1989-90 data reflect one-year (only grade 2) of time in Challenge for the pupils. The 1990-91 data reflect those pupils who had Challenge class-size reduction (1:15) in grades one (1989-90) and two (1990-91). (See Table 8.)

Table 8 about here

Although there clearly are limitations, one fairly simple way to see if Challenge systems as a group (n=17) seem to be benefitting from the treatment (i.e., 1:15) is to consider the rankings (or the aggregate rankings) of the 17 Challenge systems among all Tennessee systems (n=138). This was done for reading and for math by adding the rankings of the 17 Challenge systems (using data provided by the SEA) and then dividing by 17 to get the "average" ranking in 1989-90 (baseline) and then in subsequent years (e.g., 1990-91). Since a rank of "one" is best, a gain is achieved when the aggregate (and average) ranks become lower. With a total of 138 systems, the state average rank would be 69.

Data in Table 9 show that, on average, the Challenge systems moved up 5.3 ranks in reading and 6.6 ranks in math from 1989-90 to 1990-91. The average Challenge system (1990-91) was at 94 in reading and 79 in math, still below the state average (69).

A second procedure is to convert the district average scores to Z-scores and then to consider how the 17 Challenge system's grade-two average scores in reading and math deviate (e.g., in terms of standard deviation units) from the state average. Although the average Z-scores for reading and for math for both 1990 and 1991 TCAP results are below the state

average, the .23 and .26 standard deviation gains moved these 17 systems closer to the state mean from 1990 to 1991 testings in both reading and math (Table 10).

Tables 9 and 10 about here

Gains in rankings and in Z-score comparisons show that, on average, the second grade TCAP results are going in the desired direction; student scores are getting better as the systems move closer to the state averages. Subsequent analyses will see if the trend continues.

DISCUSSION

The power of the design and therefore the strength of the results and the confidence that one has in the findings/conclusions diminish as one moves from the experiment of STAR to the LBS field study and finally to the suggestion that application of STAR findings is helping improve student achievement in Project Challenge. On the other hand, the STAR results help in determining ways that achievement can be improved in Challenge schools and they help in understanding the changes that are occurring.

Class size reduction, as a treatment or intervention, is really an one-time event. That is. the treatment is when the student first experiences the reduction from regular (e.g., 1:28) to small (1:15); the ensuing years are a continuation, but not a separate treatment.

Challenge systems gained in the state rankings, but the magnitude of the gains was less than the demonstrated gains in STAR. Although consistent in all STAR conditions (S, R, RA), pupil assignment in STAR (random) was different from regular pupil assignment practices. Did pupil random assignment positively influence STAR results in all or in some STAR conditions? Additional analyses of the STAR database may help unravel this interesting question.

The LBS results show the continuing benefits of a pupil's participation in the small class. Post hoc analyses of important elements of schooling other than achievement (e.g., participation) suggest a small-class influence here, too. Continuing analyses through LBS will add to information provided by other longitudinal studies (e.g., Weikart, 1989; Zigler, 1992) of important social benefits of early primary and pre-primary interventions.

Since LBS shows continuing benefits in pupil achievement after small-class involvement, will small-class involvement for one or two years (rather than STAR's four years) provide a sound base to help pupils get started well in school? If so, STAR results were strongest in K and 1, suggesting that these should, at a minimum, be the years of the small-class intervention. The early primary heterogeneous classes provided by the STAR random assignment and STAR's seeming ability to help minority pupils close the achievement gap are promising areas for LBS analyses. The Ramey (1992) model may help here.

Results of STAR (the experiment) provide clear evidence of ways to improve schooling in early primary grades. Given the added needs of children entering schools in the 1990's (e.g., Hamburg, 1992; Hodgkinson, 1991) the use of small classes may become imperative for later school success. We have found a way to improve schooling; do we have the will? The STAR experiment results have held up in field research and policy conditions (e.g., LBS, Challenge) and are continuing to show added, continuous benefits. How much evidence do leaders need before they apply the findings to help improve schooling?

The progression of research from experiment (STAR) to field study (LBS) to policy (Challenge) is, of itself, an interesting approach. Table 11 shows graphically this extended emphasis on class-size issues. The consistency of results in all three approaches adds strength to the findings of each study.

Table 11 about here

Some speculation is interesting here. If small classes reduce retention in grade (STAR showed a reduction in retention in grade 1) and if there is a major "gap reduction" that may reduce the need for remediation later (in grade 1 small-class minority pupils perform more than 17% better than their large-class peers) and if the reading and math benefits occur in small classes in less time of instruction [64 minutes (small) vs 84 minutes (regular) for reading instruction], then these added benefits should be considered in addition to just the achievement results. (Analyses of some of these points are proceeding.)

Small classes are a facilitating variable. They seem to let teachers in early primary concentrate on teaching. Note that "Success for All" (Slavin et al., 1990) builds on a base of 1:15 and "Reading Recovery" is a very successful tutorial program.

Should these and similar studies be seen simply as studies in class size reduction? Perhaps they are better cast as trying to find the right class sizes to help solve Bloom's (1984) "two-sigma" problem -- trying to match the size of the instructional unit to the job to be done. The results suggest ways to move from assembly-line, industrial-age schooling to case-load, information-age learning activities. Will educators seize the initiative in the information age? It is education's time! Let's do it - Now!

References

Bloom, B. (1984, May). The search for methods of group instruction as effective as one-to-one tutoring. Educational Leadership, 4-17.

Cahen, L.S., Filby, N., McCutcheon, G., & Kyle, D.W. (1983). Class size and instruction. New York: Longman.

Education Research Service or ERS. (1978). Class size: A summary of research. Arlington, VA: Author.

Education Research Service or ERS. (1980). Class size research: A critique of recent meta-analysis. Arlington, VA: Author.

Finn, J.D. (1989, Summer). Withdrawing from school. Review of Educational Research, 59(5), 117-142.

Finn, J.D., & Achilles, C.M. (1990, Fall). Answers and questions about class size: A statewide experiment. American Educational Research Journal, 27(3), 557-577.

Finn, J.D., Achilles, C.M., Bain, H.P., Folger, J., Johnston, J., Lintz, M.N., & Word, E. (1990). Three years in a small class. Teaching and Teacher Education, 6(2), 127-136.

Finn, J.D., & Bock, R.D. (1985). MULTIVARIANCE VII user's guide. Mooresville, IN: Scientific Software, Inc.

Finn, J.D., & Cox, D. (1992, Spring). Participation and withdrawal among fourth-grade pupils. American Educational Research Journal, 29(1), 141-162.

Finn, J., Zaharias, J., Fulton, D., & Nye, B. (1989, Fall). Carry-over effects of small classes. Peabody Journal of Education, 67(1), 75-84. (Published in 1992).

Folger, J. (Ed.). (1989, Fall). Project STAR and class size policy. Peabody Journal of Education, 67(1). (Published in 1992.)

Glass, G.V., & Smith, M.L. (1978). Meta-analysis of research on the relationship of class size and achievement. San Francisco: Far West Laboratory for Educational Research and Development.

Glass, G.V., Cahen, L.S., Smith, M.L., & Filby, N.N. (1982). School class size. Research and policy. Beverly Hills: Sage Publications.

Hamburg, D.A. (1992). Today's children. New York: Time Books, Random House.

Hodgkinson, H. (1991, Sept). Reform vs reality. Phi Delta Kappan, 73(1), 8-16.

Mitchell, D.E., Beach, S.A., & Badarak, G. (1989, Fall). Modeling the relationship between achievement and class size: A re-analysis of the Tennessee Project STAR data. Peabody Journal of Education, 67(1), 34-74. (Published in 1992).

Mitchell, D.E., Carson, C., & Badarak, G. (1989, May). How changing class size affects classrooms and students. University of California, Riverside: California Educational Research Cooperative.

Mueller, D.J., Chase, C.I., & Walden, J.O. (1988). Effects of reduced class sizes in primary classes. Educational Leadership, 45, 48-50.

Nye, B., Zaharias, J., Fulton, D., Achilles, C.M., & Hooper, R. (1991 and 1992). The lasting benefits study: Technical reports (grades 4 and 5). Nashville, TN: Tennessee State University Center for Excellence.

Orlich, D.C. (1991, April). Brown v. board of education: Time for a reassessment. Phi Delta Kappan, 72(8), 631-632.

Ramey, M. (1992). (1992, April). Classroom characteristics related to ethnic achievement gap reduction. Paper at AERA, San Francisco. (Ramey is at Seattle Public Schools.)

Robinson, G.E. (1990, May). Synthesis of research on the effects of class size. Educational Leadership, 47(7), 80-90.

Robinson, G.E., & Wittebols, J.H. (1986). Class size research: A related cluster analysis for decision making. Arlington, VA: Educational Research Service, Inc.

Shapson, S.M., Wright, E.N., Eason, G., & Fitzgerald, J. (1980). An experimental study of the effects of class size. American Educational Research Journal, 17, 141-152.

Slavin, R.E. (1989). Achievement effects of substantial reductions in class size. In R.E. Slavin (Ed.), School and classroom organization (pp. 247-257). Hillsdale, NJ: Lawrence Erlbaum Associates.

Slavin, R.E. (1990, Fall). Class size and student achievement: In small better? Contemporary Education, LXII(1), 6-12.

Slavin, R.E., Madden, N.A., Karweit, N.J., Liverman, B.J., & Dolan, L. (1990, Summer). Success for all: First-year outcomes of a comprehensive plan for reforming urban education. American Education Research Journal, 27(2), 255-278.

Tomlinson, T.M. (1988). Class size and public policy: Politics and Panaceas. Washington, DC: US Department of Education, Office of Educational Research and Improvement.

Tomlinson, T.M. (1990, Fall). Class size and public policy: The plot thickens. Contemporary Education, LXII(1), 17-23.

Weikart, D.P. (1989, June). Quality preschool programs: A long-term social investment. Occasional paper Number 5. Ford Foundation Project on Social Welfare and the American Future. New York: The Ford Foundation. (28 pp.).

Word, E., Johnston, J., Bain, H., Fulton, B., Zaharias, J., Lintz, N., Achilles, C.M., Folger, J., & Breda, C. (1990). Student/teacher achievement ratio (STAR): Tennessee's K-3 class size study. Final report. Nashville, TN: Tennessee State Education Department.

Zigler, E. (1992, June 27). Head Start falls behind. New York Times, p. 15.

Table 1. STAR Kindergarten (1985) Pupils Shown by Their Distribution (%) on Selected Demographic Variables into the Three Class Types (S, R, RA).

	CLASS TYPE						
	S		R		RA		Total
Total N	1900		2194		2231		6325
% by Type (Tot)	30.0	Dif*	34.7	Dif*	35.3	Dif*	100
% Male	30.1	+.1	34.4	-.3	35.5	+.2	100
% Female	30.0	0	35.0	+.3	35.0	-.3	100
% Nonwhite	29.0	-1.0	34.5	-.2	36.5	+1.2	100
% White	30.6	+.6	34.8	+.1	34.7	-.6	101**
%Free lunch	29.2	-.8	34.2	-.5	36.6	+1.3	100
% No free lunch	30.8	+.8	35.2	+.5	34.0	-1.3	100
% Sp ed	35.6	+5.6	33.2	-1.5	31.2	-4.1	100
% No sp ed	29.9	-.1	34.7	0	35.4	+.1	100

*Difference (+, -) from "expected" distribution based on the proportion in Total. If 30.0% of students are in S, 30.1% of males in S would be +.1%. **Rounding.

Table 2. Parameters of STAR: Totals and Research Tapes, Grades K-1.

	Dist.	Sch.	Pupils	Classes (N) (%)							
				S		R		RA		Tot.	
1985-86 (K)	N	N	N	N	%	N	%	N	%	N	%
Totals	42	79	6325	127	38.7	103	31.4	98	29.9	328	100
Res Tape**	42	79	5734	127	38.7	103	31.4	98	29.9	328	100
1986-87 (Grade 1)											
Totals	42	76	7103	124	35.7	115	33.2	108	31.1	347	100
Res Tape**	42	76	5905	124	35.7	115	33.2	108	31.1	347	100

*S=1:15; R=Regular; RA=Regular with Teacher Aide.
**The research tape included pupils who met various criteria. Not all pupils had scores for all measures each year. Participation in grade one is greater than in (K) due to Tennessee not having required (K); new pupils entered and were randomly assigned.

Table 3. Average Percent of Pupils Passing BSF Reading: Grade 1, STAR.

Status	Grade	Class Type		Difference (S-R) or (S) Advantage
		Small	Reg.	
Minority	1	65.4%	48%	17.4
Non-Minority	1	69.5%	62.3%	7.2
Difference		4.1%	14.3%	

TABLE 4. Estimates of (S) Effect Sizes, Using (S) and (R & RA) ÷ 2* for White (W), Minority (M) and All Pupils, K, 1, 2 and 3, STAR, 1985-1989.

Scale	Group	Grade			
		K	1	2	3 **
SAT Tests					
Total Read	W	-	.17	.13	.17
	M	-	.37	.33	.40
	All	.18	.24	.23	.26
Total Math	W	.17	.22	.12	.16
	M	.08	.31	.35	.30
	All	.15	.27	.20	.23
BSF Tests					
BSF Read	W	-	4.8%	1.6%	4.0%
	M	-	17.3%	12.7%	9.3%
	All	-	9.6%	6.9%	7.2%
BSF Math	W	-	3.1%	1.2%	4.4%
	M	-	7.0%	9.9%	8.3%
	All	-	5.9%	4.7%	6.7%

*Effect size is difference divided by the appropriate standard deviation (for groups or totals). The BSF percents are calculated from differences of groups in percent passing. No BSF tests were given in K. Grade 2 computed on untrained teachers only (N = 273).
**Grade three was computed on Total Language Test results.

Table 5. LBS Results, Grade 4 (1989-90) and Grade 5 (1990-91) on TCAP. Summary of Class Effects Analysis Using Mean Scores of Sets.

	Set 1 Verbal		Set 2 Math/Sci		Set 3 Soc Sci/Study	
	4	5	4	5	4	5
Loc. (urban, etc.)	$p \leq .001$	N/A	$p \leq .001$	N/A	$p \leq .001$	N/A
Type (S,R,RA)	$p \leq .001$	$p \leq .01$	$p \leq .001$	$p \leq .01$	$p \leq .001$	$p \leq .01$
Loc X Type	NS	N/A	NS	N/A	NS	N/A
	(Results found in all locations equally)					

Loc. differences on all sets favoring S in the location, but major difference is due mostly to lower-performing inner-city pupils. Type differences favor S. R vs RA contrasts NS. Loc X Type class-type differences are the same in all locations.

Table 6. LBS: Grades 4 and 5. TCAP. Scaled Score Differences and the Differences in Mean Number of Domains Mastered between S and R Class Students and between RA and R Class Students. Means are tabled in Appendix B of the Technical Report (Nye et al., 1991, 1992).

Measures	1989-90 (4th)		1990-91 (5th)	
NRT	S vs R	R vs RA	S vs R	R vs RA
Total Reading	5.61	-2.23	10.53	.10
Total Language	4.99	-.73	8.21	-1.03
Total Math	4.87	-2.29	8.08	-.34
Science	5.69	-1.47	8.99	-2.66
Social Sciences	6.13	-.195	8.14	-1.31
Study Skills	10.10	-2.15	10.62	-.85
CRT (Domains Mastered)				
Language Arts	.25	-.18	.84	.07
Mathematics	.35	-.09	.68	.16

Table 7. LBS: Grades 4 and 5, 1989-90; 1990-91. TCAP. Estimates of S and RA Effect Sizes.

Measures	1989-90 (4th)		1990-91 (5th)	
NRT	S v R	R v R A	S v R	R v RA
Total Reading	.13	-.05	.22	.06
Total Language	.13	-.02	.18	-.02
Total Math	.12	-.06	.18	-.01
Science	.12	-.03	.17	-.05
Social Science	.11	-.04	.17	-.03
Study Skills	.14	-.03	.18	-.01
CRT				
Language art	.11	-.09	.34	.03
Mathematics	.16	-.04	.28	.07

Table 8. Summary Table of Students in Project Challenge (TN: 1990-93) and Years of Testing Using TCAP Tests to Analyze Challenge Successes*.

Testing Year (Date) (TCAP)	Grade-2 pupils' experience in Challenge (in years) (in years) by grade(s) at time of Testing		
Test Date	Years in Challenge	Grades of Challenge	Test Used/Grade
1990	1	grade two only	TCAP, Grade 2
1991	2	grades one and two	TCAP, Grade 2
1992	3	grades K-2	TCAP, Grade 2
1993, etc.	3	grades K-2	TCAP, Grade 2

*Challenge reduces class size (1:15) in grades K-3.

Table 9. Rankings of Challenge districts (n=17) of 138 TN School Systems Based on Grade 2 TCAP Scores (Reading and Math). (Average rank is 69).

	Reading		Mathematics	
	89-90	90-91	89-90	90-91
Sum of Ranks	1681	1591	1448	1336
÷ by 17	98.9	93.6	85.2	78.6
Difference	(+90)		(+112)	
÷ by 17	5.3 RK	5.3 RK	6.6 RK	6.6 RK

Table 10. Comparison of Challenge Systems (n=17) Average Z-Scores for Reading and Math, Grade 2, TCAP Results.

	Reading		Mathematics	
Year	89-90	90-91	89-90	90-91
Z-Score	-.75	.-52	-.34	-.08
Difference	Gain (.23)		Gain (.26)	

Table 11. Relationships of STAR, LBS and Challenge Showing Years, Grades, Measurements, etc; 1985-1993.

Study	Years	Grades	Measurement	Instruments
STAR*	1985-89	K-3 1 grade/yr	Each year & longitudinal	SAT/BSF & questionnaires
LBS*	1990-93	4 - 6		
Cognitive	1990-91	4 - 6	Each year	TCAP
Particip.	1990, ?	4	Grade 4	Questionnaire
Challenge**	1989-93	K - 3 Every year	Grade 2	TCAP

*Pupils progressed through the grades and were tested each year.
**All pupils in grades K-3 every year; tested in grade 2 only. LBS and Challenge are expected to continue.

Appendix A
DATA COLLECTION INSTRUMENTS: STAR, 1985-1989

1. Profiles: Data collected include:

   System: Enrollment, total expenditures per student, location, etc.

   School: Type, size, type of community served, special programs, etc.

   Principal: Age, sex, race, education, experience, etc..

   Teacher: Age, sex, race, education, certification, experience, career ladder level, attendance, etc.

   Aide: Age, sex, race, education, experience as an aide.

   Project Student: Age, sex, race, SES, special education programs.

   Comparison Student: Age, sex, race, and SES.

2. Stanford Early School Achievement Test (SESAT II) and other forms of SAT to measure pupil achievement in math and reading/language arts, based on national norms.

3. Self-Concept and Motivation Inventory (SCAMIN) to measure elements of academic self-concept and academic motivation.

4. Basic Skills Mastery (BSF). A curriculum-based criterion-referenced test to measure mastery of objectives in grades 1, 2, and 3.

5. Grouping Questionnaire to study how teachers regularly divide students into groups for instruction.

6. Parent/Teacher Interaction Questionnaire to determine the amount of time teachers spend interacting with parents during a school year.

7. Teacher/Problem Checklist (Cruickshank) to measure teacher perceived problems related to class size and pupil/teacher ratio.

8. Teacher Log provides a self reported use of school time (also Aide log).

9. Aide Questionnaire to obtain basic information regarding aides' supervision, job description and training.

10. Exit Interviews to obtain teacher perceptions pertinent to the project.

APPENDIX B

Primary and Extended Analyses Designs: STAR (1985-1989); LBS 1990-1991.

Sample Design:

4 Locations (Urban, rural, etc.)	(Fixed Effect)
Schools nested in Locations	(Random Effect)
Class types (S,R,RA) crossed with locations and school types	(Fixed Effect)
2 Training categories*	(Fixed)

Source Table

Source of Variation:	Error Term:
Location (L)	Schools
Training* (TR)	Schools
Type (T)	School x type
LxT	School x type
LxTR	School
TxTR	School x type
LxTxTR	School x type

		Degrees of freedom (df)	
		Ach. Meas.	Noncog. Meas.
Schools	e.g. (1986)	75	69
School x Type	e.g. (1987)	149	137
Classes within School-Types (etc.)		Etc.	

Primary Model: Measures		Matched
Achievement (Ach):	SESAT, SAT, BSF	t-tests
Noncognitive (Noncog):	SCAMIN, Attendance, Behavior, etc.	

Extended Model: Measures:		
Sex (or Race, or SES)	Ave. Diff Scores on Ach.	Multivariate
Sex (or Race, or SES) Training*	Ave. Diff. Scores on Noncog.	Models

Two planned contrasts: S class mean vs means of all R and RA; S vs (R + RA ÷ 2)
RA class mean vs R class mean.

Each effect tested holding constant earlier effects in order of elimination. TR and T each tested as last main effect; LxTR and LxT each tested as last two-way interaction. Analysis of BSF done with "log-odds index."

*For grades 2 and 3, a random subset of schools was chosen to study the effects, if any, of teacher training (TR) on pupil outcomes. Although not discussed in detail here, the training used had no significant effect.

Results appear in various other articles and reports.

APPENDIX C
Analysis of Variance for Cognitive Outcomes, STAR, Grades K-3.
Sig. Levels p≤.05 or Greater are Tabled.

Effect/[a] Grade		Reading			Mathematics		
		Multi-variate [b]	SAT Read	BSF Read	Multi-variate [b]	SAT Math	BSF Math
Location (L)	K		.02			.05	
	1	.01	.06		.05		
	2	.001	.001	.001		.001	.001
	3	.001	.001	.001	.001	.001	.001
Race(R)	1	.001	.001	.001	.001	.001	.001
	2	.001	.001	.001	.001	.001	.001
Type(T)	K		.001			.02	
	1	.001	.001	.001	.001	.001	.05
	2	.001	.001	.05	.001	.001	.05
	3	.001	.001	.001	.001	.001	.001
SES	K		.001			.02	
Loc X Race	1	.05		.05			
Loc X Type	K - 3	All N/S. The class-size effect is found equally in all locations -- Inner City, Suburban, Urban and Rural schools. (Tabled as important.)					
Race X Type	1	.05	.05	.01			
LxRxT	1			.05			.01
LxTRxT	2	.05	.01	.05	.05	.05	.01

NOTE: Only statistically significant (≤.05)results are shown. [a] The nonorthagonal design required tests in several orders (Finn and Bock, 1985). Results were obtained as follows: each main effect was tested eliminating both other main effects; loc x race tested eliminating main effects and loc x type; loc x type tested eliminating main effects and loc x race; race x type tested eliminating main effects and other two-way interactions, and loc x race x type tested eliminating all else (Finn and Achilles, 1990). [b] Obtained from F-approximation from Wilks' likelihood ratio. Essentially, no statistically significant differences were obtained on the self-concept and/or motivation (SCAMIN) measures. No training main effect, or training-by-type interaction. Trained and untrained teachers did equally well across all class types and the (S) advantage (and absence of Aide effect) is found equally in all four locations for trained and untrained teachers.

(S) advantage and all effects found for total class generally apply equally to white and nonwhite pupils, especially in grade 2. The race difference was statistically significant for all measures and multivariate sets, but not for most interactions (LxR, TRxR, TxR, LxT,R, or TRxTxR). (S) significantly better than (R,RA) on all tests; no R vs RA tests significant.

Results appears in other articles and reports.